The Sufficiency of Reinforcing Problem Solutions for Transition to the Formal Stage

Michael Lamport Commons and Martin N. Davidson
Harvard University

A racially and socioeconomically integrated population of fifth and sixth grade students was repeatedly presented with Formal Stage 11 single cause problems. With problem presentation alone and with problem presentation and feedback, no significant change occurred. The transition to the Formal Stage accelerated significantly across trials for participants whose correct answers were reinforced, supporting the sufficiency of reinforcement, $r = .77$, $p \leq .00005$. This result adds further support to the claim that reinforcement is sufficient for stage change and better than just practice or feedback in producing stage change.

Keywords: equilibration, Formal Stage 11, Inhelder and Piaget, reward/reinforcement versus feedback, transition

The world of stage and stage change has largely been dominated by mentalistic notions of stage. Maybe because of this, it has avoided any evolutionary basis for why stage change would be so responsive to reinforcement of next stage behavior. Here we address what methods help individuals develop next stage behavior including problem solving skills that are more advanced in terms of hierarchical complexity.

In beginning to answer this broad question, we examined a several forms of pedagogy that may help fifth and sixth grade students acquire problem solving performance at the Formal Stage 11. Performing at the Formal Stage 11, as used here, requires solving multiple variable problems containing a single causal variable. This conception of Formal Stage 11 comes from Commons' Model of Hierarchical Complexity (MHC; Commons, Trudeau, Stein, Richards, & Krause, 1998; Commons, Gane-McCalla, Barker, & Li, 2014; Commons & Miller, 1998; Commons & Richards, 1984a, 1984b). In this model, the stage of a performance is equated to the order of the hierarchical complexity of the tasks that the performance successfully addresses. *Order of hierarchical complexity* is measured by the number of recursions that the coordinating actions must perform on a set of primary elements. Actions at a higher order of hierarchical complexity (a) are defined in terms of the lower order actions, (b) organize and transform the next lower order actions, and (c) produce nonarbitrary organizations of these lower order actions that solve new, more complex tasks. Solving problems at the formal-operational stage, as this study asked its participants to do, is one of the basic elements of scientific reasoning (Inhelder & Piaget, 1958). Generally acknowledged as an important skill, scientific reasoning often greatly aids individuals in their efforts to successfully solve prob-

lems in school and in everyday life. Because many adolescents and adults cannot accomplish tasks of this Formal Stage 11 of hierarchical complexity (Neimark, 1975), understanding the factors that assist in the attainment of Formal Stage 11 is crucial.

Many different methods could be used to promote stage transition to the Formal Stage 11. One method examined by other studies is to support and help the student through the actual problem solving process. Using this technique, adults could provide examples (Kitchener & Fischer, 1990) or try to guide or prompt the correct response from students (Brown & Campione, 1990; Brown & Palincsar, 1989). This study does not examine these types of methods because other investigations have already shown that support techniques lead to the acquisition of higher stage behavior (Fischer, personal communication, 1990; Kuhn, 1990; Vygotsky, 1962, 1966, 1978). Putting this method aside, this study instead focused on interventions where different types of outcomes were administered after the students came up with a solution to a problem on their own. This method is highly different than support techniques, which intervene throughout the problem solving process. Perhaps this type of intervention could also be sufficient in bringing about performance at the Formal Stage 11.

Three different types of these interactional environments where students solve a problem without instruction and then receive different consequences include (a) repeated practice of the problems requiring Formal Stage 11 performance, (b) the Piagetian notion of immersion with the outcome of explicit pointing out of contradiction, and (c) the use of explicit reinforcement outcomes for correct answers. The efficacy of these three consequences in producing stage change to the Formal Stage 11 was investigated in this study.

Under the first of these different methods, participants solved a Formal Stage 11 problem on their own and were given no feedback. Instead, they were only exposed to more problems. Although used primarily as a control for the other two methods, the effectiveness of mere practice without feedback in helping students acquire a higher stage performance deserves investigation as well. Some studies have shown that presenting similar problems repeatedly during training without any other instruc-

tional techniques like feedback as to whether their solutions are correct or reinforcement have induced improvements in Formal Stage 11 performance. Commons and Richards (1980), for example, found that when 13- to 18-year-old adolescents from the upper-middle class suburbs of Boston were repeatedly administered problems without any other kind of helpful intervention, their performance improved. This method is probably not successful for every person, however, because of the fact that many people have probably been exposed to Formal Stage 11 problems over their lifetimes, yet many adolescents and adults still cannot solve these problems (Commons, Miller, & Kuhn, 1982; Neimark, 1975).

If participants tend not to attain a higher stage of performance with practice alone, perhaps practice along with feedback would help this change take place. This technique for the acquisition of higher stage performance is somewhat Piagetian in that it posits that children will be motivated to restructure their knowledge when they encounter experiences that conflict with their predictions. In receiving feedback that his solution is incorrect, the student's state of equilibrium is disturbed. Equilibration (Piaget, 1971) is the process by which an organism regulates and maintains a balance between the complementary processes of assimilation and accommodation. According to Piaget, development occurs in a progression through a sequence of stages in response to disequilibration. At each stage, a person obtains a momentary equilibrium. When confronted with a higher stage problem, the person is thrown out of equilibrium when he or she uses previous stage strategies and finds evidence (in this case verbal feedback) that the predicted outcome (in this case a correct answer) does not occur. This discovery learning (see Table 1) situation may

Table 1

Comparison of Motivation Conditions Associated With Discovery and Directed Learning

Motivation condition	Type of learning	
	Discovery	Directed
No feedback	Not possible	Possible
Feedback	Possible	Possible
Feedback somewhat reinforcing	Possible	Possible
Clear reinforcement	Not usual	Possible

then compel the participants to resolve this apparent contradiction. To reequilibrate, they figure out how to perform the higher order task yet still use some of their previous knowledge to come to this resolution.

Other studies have tested the efficacy of this feedback training, and their results have been quite mixed. Bredderman (1973) found that "for some children the repeated posing of a question left unanswered," in addition to verbal feedback as to the efficacy of solutions, "was sufficient inducement for a cognitive reorganization." Although we are not approaching this study of stage change from a cognitive-developmental angle, in the case of Bredderman's study a cognitive reorganization leads to a higher stage performance, which is what the conditions of this study are seeking to induce. Davidson, on the other hand, found a complete lack of improvement with repeated presentation and experimenter-feedback methods among a socioeconomically and racially mixed population of 11- and 12-year-old subjects (Davidson, 1983; Davidson & Commons, 1983). Thus, such verbal feedback may aid some children in transition to Formal Stage 11 performance and may be ineffective in inducing this behavior in other children.

One factor that could account for the varied results of these studies is that Piagetian-like models assume that the participant values both detecting (Flavell, 1971; Kessen, 1971; Langer, 1969) and resolving inconsistency. This assumed motivation was one of the aspects that made Piaget's theory so novel at the time it was proposed. Two of the most influential theories of the time- Freud's personality psychology and Hull's experimental psychology—both stressed that other than fulfilling sexual or hunger drives, individuals sought to avoid overwhelming stimulation from the outside world (Ginsburg & Opper, 1988). In Piaget's theory, individuals not only seek intellectual stimulation but take this drive so far as to resolve any conflicting intellectual stimulation they may encounter. Perhaps more like Freud and Hull's models, some individuals may not be interested in undergoing the intellectual strain and stimulation implicit in any transition to a higher stage of performance. Because different children are raised with different values in different cultures and subcultures, maybe some students do not value the outcome of arriving at a correct answer in and of itself as much as other students, and are thus not compelled to figure out the solution to the higher stage problems. It could also be the case that some individuals just get discouraged and give up if they receive too much negative feedback, rather than being further determined to solve the problem at hand.

Piaget himself might account for the varied results of these studies by pointing to the fact that the study participants did not choose the problems or tasks themselves. Because these studies involved posing problems to the subjects, the reinforcement inherent in posing one's own questions is lost. Because the "posing of problems" paradigm is similar to that which occurs in schools, however, it is important to investigate methods which might help substitute for the lack of motivation to learn whether some students might progress under other teaching methods.

To investigate one technique that may provide the motivation that these students may require, the interactional environment of external reinforcement was also included in the study. Reinforcement is the process of presenting consequences following behavior that increase or maintain the behavior that the consequences follow (Skinner, 1938). McCann and Prentice (1981) compared the effects on moral development of four conditions: direct reinforcement (DR), cognitive disequilibrium (CD), combination (COM) of both DR and CD, or control in 40 elementary schoolchildren (Mean age = 7.8). DR, CD, and COM groups were trained on moral judgment items adapted from Piaget's (1965) studies of intentionality. Following training, moral judgment was assessed through administration to all groups of an immediate posttest, switch items (a variant of the training items), a generalization test, and a 2-week follow-up on the original training items. The ANOVAs indicated that DR was most effective in promoting change in moral judgment initially but that effects of CD were more enduring and generalized. The COM condition, perhaps because of its complexity, was relatively ineffective. Conrath's (1988) study on discouraged or defeated learners from desperately poor, violent, or drug-addicted families showed that this student group benefits educationally and psychologically from external reinforcement. Perhaps external reinforcement of correct problem solutions will be most effective

at bringing about a universal acquisition of Formal Stage 11 performance among a diverse participant pool. Examining reinforcement, feedback, and practice methods, this study is trying to answer that very question. Which of the three methods is most effective method for helping a diverse group of students acquire Formal Stage 11 performance? A systematic examination of the efficacy of different consequences after students solve a problem on their own was undertaken in order to provide a thorough response.

Method

Participant Selection

In any study on stage change, participants must be selected carefully to ensure that they do not possess any prior understanding of the particular developmental concept for which they are being trained and tested. In this study, it was necessary to select young participants who did not perform at the Formal Stage 11 stage to see what level of intervention was sufficient for inducing actual rather than apparent stage change. Apparent stage change occurs when in the course of the study, the subject transfers preexisting Formal Stage 11 proficiency from some other task or domain to solve the task administered under the conditions of the study. This apparent stage change took place in an earlier study done by Commons, Miller, and Kuhn (1982) on formal stage performance in college students. Consistently improving by the fourth trial of the formal stage problem, the college students' performances were most likely attributable to a warm-up effect and subsequent transfer of training to the new domain of the study's problem rather than an actual stage change where participants learned how to do a Formal Stage 11 task for the first time.

To find an age group of similarly trained students that did not yet possess any experience successfully solving Formal Stage 11 problems, we based our subject selections on the results of studies by Kuhn (Kuhn & Angelev, 1976; Kuhn & Ho, 1977; Kuhn, Ho, & Adams, 1979). They found little performance above the Concrete Stage 9 as compared with the Formal Stage 11 up to and including grade six. Most of the students in sixth grade only perform prefor-

mally at the Abstract Stage 10 and Concrete Stage 9. In fact, most students in their studies did not exhibit Formal Stage 11 performance until grade eight. Thus, in light of these findings, we formed our subject pool from elementary school students in the fifth and sixth grades. We did this to avoid any transfer of training effects and to focus on what are sufficient causes rather than apparent causes for stage change.

Over a period of six years, all children in two fifth- and two sixth-grade classes were invited to participate in the study. The 134 children who returned their parental-permission slips participated. The participants attended three different racially integrated elementary schools in the same city. Most of the families had lower to lower-middle class socioeconomic status, and approximately one half were racial or ethnic minorities.

Experimental Design

To assess the effects of problem presentation (practice), feedback, and reinforcement, separate groups of these fifth and sixth grade children were exposed to five different amounts of intervention as they attempted to solve a causality problem. This level of intervention is the B in the ABA single-case experimental design that this study adapted from Hersen and Barlow (1976); Hersen and Barlow (1976). Here, a pretest was given, a level of intervention was administered, and then a posttest of the same task was administered to assess whether the intervention induced any stage change in the participants.

Receiving no intervention (B), participants in the Pretest-Posttest Group (G-1) served as a control for all the other groups and received only the pre- and posttests. Participants in the other conditions received pre- and posttests and participated in the intervention with the following differences. In the Intervention Group (G-2) participants solved additional problems but did not receive feedback as to whether their responses were correct or incorrect. In the Feedback Group (G-3), participants received verbal feedback as to whether their responses were correct or incorrect. Students were randomly assigned to these three groups for all six years of the study.

Because these interventions failed to facilitate the students' Formal Stage 11 stage performance during the first three years, a new intervention group, the Reinforcement Group, was added for the remaining three years of the study. Participants in the Reinforcement Group (G-4) received verbal feedback and each correct response was immediately reinforced with points. Although each subject answered questions individually like the other participants in the study, the points they received for correct answers were added to a group total. The points were part of a group game in which teams were organized to compete against each other. Each member of the team that scored the most points at the end of the entire problem sequence received a prize. Prizes were chosen by each individual child from a list of prizes generated by all of the children participating in the portion of the study.

This method differs from studies that reinforced overall performance at the end of a series of tasks in that it divides children into groups to make the points more reinforcing (Gruen, 1965; Siegler & Atlas, 1976; Siegler, Liebert, & Liebert, 1973). Johnson, Maruyama, Johnson, Nelson, and Skon (1981) found that this method of having participants work at an individual task toward a group reinforcement to be the most effective method among the number they reviewed. In two studies, Commons and Goodman (2008) and Fargo, Behrns, Goodman, and Commons (1971) showed that this method is particularly congruent with values of low-income African American families, who were a large proportion of the participants in this study. Thus, competitive groups were added to the experimental design in order to make sure that all of the students in the reinforcement group would be most thoroughly reinforced for their correct answers and Formal Stage 11 performance.

Participants in the Construction Group (G-5) were exposed to these same conditions of reinforcement, except that they were asked to perform an additional task of constructing and then solving a portion of the problem. This Construction Group was added in the final year of the study to investigate whether increasing the child's active involvement in the task facilitated stage transition. A summary of the conditions for each group is shown in Table 2.

Pretest and Posttest Procedure and Materials

Written versions of the plant problem (Kuhn & Brannock, 1977; Kuhn & Brannock, 1977), pendulum problem (Inhelder & Piaget, 1958), and paint problem (Commons & Rodriguez, 1987) served as pretests (A) and posttests (A′). The administration of each pre- and posttest problem type was counterbalanced within a subject and randomized across participants and groups. Participants in the four intervention conditions were then exposed to at least 16 problem presentations (B) over a course of approximately five months. Note that the Paint Problem was derived from the Pendulum (Inhelder & Piaget, 1958), Plant (Kuhn & Brannock, 1977; Linn, Chen, & Thier, 1976), and Laundry Problems (Commons, Miller, & Kuhn, 1982). The plant problem asked participants to determine which variable (large or little pot, a lot or little water, reddish or yellowish dirt, blue or red plant food) was causal for the plant outcome of healthy or sick. The paint problem contained the following possible causal variables: water or oil-based paint, wet or dry cloth, coarse or fine sandpaper, paint brush or paint roller. The outcome was whether the paint would be smooth or cracked. The Pendulum problem contained the following variables:

Table 2
Problem Presentations Groups

Group	n	Pretest/Posttest	Intervention	Verbal feedback	Valued reinforcement	Problem construction
G-1	61	X				
G-2	26	X	X			
G-3	9	X	X	X		
G-4	33	X	X	X	X	
G-5	4	X	X	X	X	X

heavy or light weight, short or long string, weight started high or low, weight pushed hard or softly. The outcome was whether the pendulum swung across the center line more or less often.

Intervention Procedure and Materials

Each intervention problem presentation consisted of one of eight versions of Commons, Miller, and Kuhn's (1982) laundry problem. Details are presented below. Each of the 16 intervention problems was presented in the form of a different kind of stain (D. Kuhn, personal communication, September 1980). Therefore, although each configuration of variables was repeated once, this was not apparent until the problem had been solved. The solution of the laundry problem required the isolation of a causal variable and the rejection of the noncausal variables as having an effect on the outcome. This is a task requiring Formal Stage 11 performance.

The laundry problem was chosen as the intervention problem in this study for a number of reasons. First, washing is a task that occurs in all cultures. The outcomes associated with washing may then have more relevance to members of other cultures than would a scientific or physical task. The participants in this experiment were from highly diverse backgrounds. The use of the laundry problem instead of a problem like Inhelder and Piaget's pendulum problem might have canceled out some possible cultural bias these more scientific problems could have contained. Second, the variables are dichotomous, having discrete values (e.g., liquid or powdered soap). Additionally, like the causal variables in the plant and pendulum problems, the causal variable in the wash problem differed with each version, underscoring the hypothetical nature of these problems. No single variable was likely to become more pronounced than any other. This variability also ensured that the stain and the causal variable for each trial did not usually match real-world experience, further establishing the hypothetical nature of the laundry problem. Finally, the nature of the laundry problem lends itself to a schema in which advancements in performing toward a Formal Stage 11 lead to a higher percentage of correct answers.

Each laundry problem comprises 16 episodes, each including four independent variables (soap type, water temperature, bleach brand, and booster color) followed by an outcome variable (cloth cleanliness). The first six of these episodes are referred to as *Informational Episodes*. These episodes were used to provide the students with enough information on the various interactions between the cleaning product and the stain so that the causal variable for the correct (clean) outcome could be determined. The labeled bottles contained the actual cleaning agents. The six-inch-square cloths were either clean or stained. To provide the students with this information, each episode showed the effects of four combinations of pairs of washing agents: water, soap, booster, and bleach on stained cloths. Three of the Informational Episodes resulted in a cloth outcome of clean and three in a cloth outcome of dirty. Only one pair of ingredients actually cleaned the cloth; the others were randomly associated with the outcome.

The next 10 of these episodes are referred to as *Prediction Episodes*. Participants were asked to use the information from the Informational Episodes to determine the cloth outcome in each of the Prediction Episodes. For each episode, an experimenter asked which outcome the subject thought would occur if the cloth were washed in the given combination of cleaning agents. The experimenter then probed for the subject's reasoning as to why the cloth would come out that way. Such probes included questioning as to which variable or combination of variables (i.e., soap; or soap and bleach; or soap, bleach and booster; etc.) was responsible for how the cloth came out. Furthermore, the probes asked whether or not one variable made a difference in the outcome. Throughout the prediction episodes the informational episodes were displayed to enable participants to refer back to them and to avoid confounding the isolation of variables task with a memory task.

Most participants took the posttest after their 10 prediction episodes. Participants in the Construction Group (G-5), however, undertook an additional task in the laundry problem and constructed their own prediction episodes. They were asked to choose a combination of variables and indicate what they thought the corresponding outcome should

be. For example, they might choose hot water, powered soap, brand A bleach, and pink booster and then predict clean. If hot water produced clean, they were told "correct," thus receiving feedback on the correctness to their own, subject-chosen episodes. This additional step in the experimental design provided an additional way to investigate the question that this study sought to solve: How do different levels of intervention affect an individual's performance solving tasks of a new, more hierarchically complex order?

Note that the laundry problem is particularly appropriate for this kind of study, for the following reasons. Inhelder and Piaget (1969) suggested that because stage change is sequential, each successive phase in development is most probable given the results of the preceding phase. They reasoned that behavior peculiar to the next stage increases in probability during the completion of the previous stage, although the mechanism for this was left open in their discussion. A task analysis of the Laundry Problem (Richard, Unger, & Commons, 1988) illustrates that participants attended to fewer variables during the concrete-to-abstract stage transition than during the transition to the formal stage. The feature of the increase in correct responses during stage transition is built into this problem, which allows for reinforcement to increasingly be associated with correct answers. In this way, even small advances may lead to more reinforcing situations. The general subject of the problem, its random associations between variables, and its ability to reinforce performance that works toward the Formal Stage 11 all made the laundry problem highly appropriate for this study. Finally, the psychophysical properties of the Laundry task sequence are also well known. The stages are equally spaced ordinals (Commons, Li, Richardson, Gane-McCalla, Barker, & Tuladhar, 2014).

Results

Whether reinforcement, rather than the other levels of intervention, will be sufficient for stage change will be measured in four ways. First, the relationship between the amount of intervention and proportion of participants that ended up performing at the For-

mal Stage 11 stage was examined. Second, changes across repeated presentations of the laundry problem were also analyzed across the groups. Terminal performance was compared with initial performance to examine the relative efficacy of reinforcement in changing stage of performance versus the other variable in the study, which was simply performance over repeated presentations of the experiment. Third, the different rates of acquisition of Formal Stage 11 performance and how reinforcement positively affected this rate are presented and discussed. Finally, transfer of performance to other causal reasoning problems such as the paint, pendulum, and plant problems is also considered.

Intervention Amount and Proportion of Formal Stage 11 Final Performances

The relationship between the amount of intervention and proportion of participants that ended up performing at the Formal Stage 11 provides the most general evidence for the sufficiency of reinforcement in facilitating stage change. To examine the role of reinforcement and increased amounts of intervention on terminal performance, a regression analysis was carried out. Participants were considered to have final Formal Stage 11 performance if they had 96% or more correct of the last 30 predictions or of the posttest predictions (for the Pretest–Posttest group (G-1). Although there were no significant differences in Formal Stage 11 performance among participants at the beginning of the experiment, there was a very large difference by the end. The probit transformation linearized the proportion of individual participants showing Formal Stage 11 performance (Crunch Software Corporation, 1987). The r between type of intervention (group) and proportion of participants whose final performance was formal stage was .987, $p < .013$. Furthermore, looking at the degree of Formal Stage 11 performance, the other groups' levels of formal stage 11 performance, in terms of proportions obtaining that stage, were: Pretest–Posttest Group (G-1) $p = .02$; Intervention (G-2) $p = .13$, Feedback (G-3) $p = .22$. No other group approached the proportion of the Reinforcement Group (G-4) $p = .55$. These results support the conclusion that reinforcement is a sufficient condition for stage change.

Formal Stage 11 Performance of Differentially Motivated Participants Across Trials

To further investigate the effect of reinforcement and different levels of intervention on stage change, the students' performances on each trial of the study were examined. This analysis is based on signal detection theory which utilizes choice theory (Munsinger, 1970; Swets & Green, 1961; Swets, Tanner, & Birdsall, 1961) to analyze the data (Kantrowitz, Buhlman, & Commons, 1985; see also Kantrowitz, Buhlman, & Commons, 1985; Rodriguez, Buhlman, Kantrowitz, & Commons, 1986; Commons & Richards, 1984a, 1984b). Signal-detection theory provides a method for determining how well a signal is being detected by participants (Commons, Kantrowitz, Buhlman, Ellis, & Grotzer, 1984). In this study, then, such an analysis was used to calculate the participants' sensitivity to the causal relation (non-normal d') which was the measure of how proficient the participants were at detecting what causes a desired outcome (cloth came out clean) on each trial.

In ascertaining this proficiency, choice theory was used to classify participants' responses as to whether they were hits (assertions that the cloth came out clean when it did), false alarms (assertions that the cloth came out clean when it did not), correct rejections (assertions that the cloth came out dirty when it did), or misses (assertions that the cloth came out dirty when it did not) as shown in Table 3. Because the reinforcement given to students in the study was symmetrical, no bias was expected or found for saying clean versus saying dirty.

Table 3
Choice Analysis of Task Performances in the Wash Problem on the Red Lipstick Stain[a]

	Subject assertion	
Stimulus	Clean	Dirty
Liquid soap	Hit	Miss
Powder soap	False alarm	Correct rejection

[a] Liquid soap is causal for the outcome of a clean cloth in the Red Lipstick Stain version. Powder soap is causal for the outcome of dirty.

From choice theory, the probability of making hits and false alarms was calculated to be the following:

$$p(Hits) = \#Hits/(\#Hits + \#Misses) \quad (1)$$

$$p(False\ Alarms)$$
$$= \#False\ Alarms/(\#False\ Alarms$$
$$+ \#Correct\ Rejections) \quad (2)$$

Then non-normal d' was found from these probabilities using signal detection theory (Commons & Richards, 1984b; Kantrowitz, Buhlman, & Commons, 1985; Munsinger, 1970) as follows:

$$Non\text{-}normal\ d' = p(Hits) - p(False\ Alarms) \quad (3)$$

Here, non-normal d' represents the Formal Stage performance of detecting which causal relation holds in a particular problem and thereby predicting the correct outcome of a given episode. The causal relation is the stimulus to be discriminated from the noncausal noise. Non-normal d' of 1.00 indicates the perfect detection of Formal Stage 11 relationships whereas a non-normal d' of 0 indicates performance at the chance level. A non-normal d' of -1.00 indicates perfect detection of a causal relationship, but attributing the positive outcome to the wrong value of the variable. Such non-normal d's of -1.00 may reflect confusion about the problem or planned choices that are opposite from the one required by the problem. Non-normal d' was used in the data analysis for this study instead of d' because as D. M. Green suggests (personal communication, November 7, 1985) non-normal d' is a preferable measure over d' when averaging across participants.

The non-normal d (the sensitivity to a Formal Stage 11 single causal relation of each subject) was found for each of the 16 trials, indicating how close to Formal Stage 11 (Commons, Trudeau, et al., 1998) the participants' performances were at each presentation. The results of this data analysis are shown in Table 4. The mean non-normal d' and the standard deviations across participants in a group were derived from each subject's 10 predictions and the number of participants in that group as shown in Table 4.

Table 4
Mean Sensitivity Across Participants by Trial Number Within Each Group

Presentation (Group)	Trial number															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Group 2 (Intervention)																
M	.26	.00	.26	.12	.15	.32	.32	.13	.29	.24	.22	.34	.27	.25	.20	.23
SD	.45	.40	.51	.50	.43	.48	.39	.51	.52	.45	.53	.49	.63	.49	.58	.53
Group 3 (Feedback)																
M	.41	.22	−.13	.35	.24	.29	.36	.31	.38	.00	.64	.45	.36	.24	.30	.29
SD	.53	.41	.64	.64	.56	.65	.52	.60	.62	.53	.58	.67	.48	.64	.42	.54
Group 4 (Reinforcement)																
M	.38	.44	.45	.69	.49	.58	.64	.62	.87	.72	.79	.75	.57	.69	.83	.81
SD	.48	.52	.54	.38	.65	.49	.45	.47	.29	.41	.38	.46	.58	.49	.35	.31
Group 5 (Construction)																
M	.50	.42	.50	.45	.26	.46	.32	.53	.72	.15	.56	.25	.37	.40	.19	.55
SD	.38	.41	.67	.71	.51	.43	.51	.45	.44	.58	.63	.70	.62	.49	.55	.30

The means ranged from .00 to .87, the standard deviations from .30 to .71.

The results show the positive influence of reinforcement and then repeated intervention on student performance. The reinforcement group exhibited the most profound improvement in Formal Stage 11 problem solving, increasing in Formal Stage 11 performance from a non-normal d' of .38 to one of .81. When differences in initial performance at the outset of the intervention period are taken into account, reinforcement continues to be most highly correlated to Formal Stage 11 performance over the trials. To control for the difference in initial performance, the average of each subject's score on the first two presentations was entered as a covariate. This average was used, rather than either the pretest or the first trial alone, because the average would not confound issues of transfer (from the Plant, Pendulum, or Paint Problems) with warm-up effects.

The effect of both the type of intervention as defined in Table 2 (Group), and of the trial on student performance was further explored through a repeated measures analysis of variance (see Table 5). Reinforcement was again found to produce stage change even after correcting for initial differences in performance by using the first two trials as a covariate. Through this data analysis, it emerges that both group and trial were highly significant as determinants for student performance. The outcome that trial was also a significant predictor of student performance was expected, as individuals often get better at a task with repeated practice and fa-

miliarity. To examine further the less obvious question raised by this analysis of how the different groups contributed to the overall significance of groups as a predictor of student performance across trials, a multiple *t* test and corresponding Bonferroni post hoc analysis was performed. This analysis was performed using Crunch Software (Crunch Software Corporation, 1987) following Miller (1981) and Rosenthal and Rosnow (1985). The *t* test was $\Pr(F(1, df) > t_{ij}^2) = \Pr(F(1, 3) > t_{ij}^2) = p_{ij}$. The Bonferroni test was $[k(k - 1)/2]p_{ij} = 6p_{ij}$.

Note that the capitalized *D* is used here to show the mean d' across trials. Table 6 shows that Reinforcement Group (G-4) participants had a significantly higher mean rate of detection of the causal relation across Trials 3–16 of the Formal Stage 11 problem, $D = .680$ than did the

Table 5
Effect of Group on Mean Sensitivity to Causal Relation Across Trials Average Performance on Trials 1 and 2 as a Covariate

Source of variance	SS	df	MS	F
Between participants				
Covariate	32.75	1	32.75	25.38***
Group (A)	28.30	3	9.43	7.31**
Within Participants				
Trial number (B)	6.03	13	0.46	3.07**
Interaction	8.05	39	0.21	1.37*

Note. Non-normal $d' = p(\text{hits}) - p(\text{false alarms})$ represents how well causal variables were detected (formal Stage 11) for each group across the 16 trials.
* $p < .0690$ (ns). ** $p < .0002$. *** $p < .00005$.

Table 6
Post Hoc Comparisons Between Mean Group Sensitivity on Trials 3–16

Comparisons	*t* test *p* <	Bonferroni *p* <
Feedback (G-3) > Intervention (G-2)	ns	ns
Reinforcement (G-4) > Intervention (G-2)	.0000	.0000
Reinforcement (G-4) > Feedback (G-3)	.0011	.0066
Reinforcement (G-4) > Construction (G-5)	ns	ns
Construction (G-5) > Intervention (G-2)	ns	ns
Construction (G-5) > Feedback (G-3)	ns	ns

Note. Here we look at what contributed to the significance of groups: which groups contributed the most and which the least.

Intervention Group (G-2) participants, $D = .240$, $p < .0000$. Reinforcement Group (G-4) participants' mean rate of detection was also significantly higher than that of the Feedback Group (G-3) participants, $D = .292$, $p < .0011$ and higher than that of the Problem Construction Group (G-5), $D = .407$, although in this latter case the *p* was not significant but at a level approaching significance. Because this group (G-5) included only four participants, a significant difference might have been found with more participants. Because the mean performance of the reinforcement group, as compared with the feedback and construction group, was significantly different, whereas the difference in mean performance between the feedback, intervention, and construction groups was not significant, the superior performance of the reinforcement group students again emerges as the most highly influential factor in the finding that group was a strong predictor on student performance.

Rate of Acquisition of Formal Stage 11 Performance

Moving a step beyond examining Formal Stage 11 performance at each trial, we next found and compared the different rates of acquisition of Formal Stage 11 performance among the varied intervention groups. To carry out this analysis, a regression of non-normal *d'* versus trial was performed (see Table 7). For each group, trial was used to predict the mean non-normal *d'* performance per trial across participants. Because a trial consisted of 10 predictions, the actual *n* included in the analysis was far greater than the 16 number of trials. Further, *n* is augmented by the number of participants

and the 10 predictions per trial. An analysis within each group indicated that there was no significant difference between performance on presentations across trials for participants in the Intervention Group (G-2), Feedback Group (G-3), and Problem Construction Group (G-5). There was a significant increase in performance across trials for participants whose correct predictions were reinforced (G-4), $r = .77$, $p \leq .00005$. This result adds further support to the claim that reinforcement is sufficient for stage change and better than just practice or feedback in producing stage change.

Transfer of Training and Reinforcement: Pretest–Posttest Analysis

Finally, an analysis of covariance of Posttest scores was performed to see whether (a) reinforcing correct predictions made a significant difference, and (b) effect of reinforcement versus the other interventions transferred to dissimilar Formal Stage 11 causality problems. The pretest scores were used as a measure of initial level of Formal Stage 11 as shown in Table 8. These pretest scores and the problem type on the Posttest were both used as covariates. The pendulum problem, which the students were presented with after the laundry problem, was significantly more difficult than the plant problem ($p < .05$). As shown in Table 8, mean performance on the Posttest for this problem was highest for the Reinforcement group (G-4), $M = 0.44$, $SD = .22$, and next highest for the Intervention alone group (G-2), $M = .16$, $S.D. = .26$. The other groups did much worse. The analysis of covariance in Table 9 shows that the within-group effect of the Posttest was significant,

Table 7

Mean Formal Stage 11 Performance as Measured by Non Normal d′ Across Trials 1–16 Within Each Group

Group	n	Mean	SD	Correlation	$p <$	Non-normal $d′$ (D) regressed on trial
Intervention (G-2)	16	0.21	0.08	0.30	ns	$D = .164 + .005 *$ TRIAL
Feedback (G-3)	16	0.30	0.18	0.22	ns	$D = .226 + .008 *$ TRIAL
Reinforcement (G-4)	16	0.65	0.15	0.77	.0005	$D = .440 + .024 *$ TRIAL
Construction (G-5)	16	0.41	0.15	−0.18	ns	$D = .462 − 0.006 *$ TRIAL
Trial	16	8.5	4.76			

$F(4, 126) = 7.30$, $p \leq .00005$. In the skill of transfer of training, the Reinforcement Group (G-4) performed significantly better than all other groups.

As additional evidence of this result, both the mean change scores in Table 8 and the Multiple t post hoc analysis, as shown in Table 10, indicate that participants in the Reinforcement Group (G-4) performed significantly more often at the formal stage, $D = .44$, than participants in Pretest-Posttest Group (G-1), $D = −0.08$, $p < .0000$, and those in the Intervention Group (G-2), $D = 0.16$, $p < .0234$. Participants in the Reinforcement Group (G-4) also performed more often at the formal stage than those in the Feedback Group (G-3), $D = 0.05$, $p < .0261$, and those in the Problem Construction Group (G-5), $D = −0.30$, $p < .0033$. Therefore, transfer of training did occur in conditions that included reinforcement, rather in the intervention conditions that did not include reinforcement, which shows that reinforcement is sufficient for initiating the application of training from one Formal Stage 11 task to another.

The superiority of reinforcement in helping participants acquire Formal Stage 11 was determined through analyses of the relationship between the amount of intervention and proportion of participants that ended up performing at the Formal Stage 11, changes in group performance across repeated presentations of the laundry problem, the different rates of acquisition of Formal Stage 11 performance among the different intervention groups, and the likelihood of the different groups to transfer of Formal Stage 11 performance to other causal reasoning problems. There could be some criticism, however, that these findings are biased because the different problem types or student cohorts from different years that were used in the study. To address worries about the varied problem types that were used, an analysis of covariance of pretest data, which was part of the last section on transfer of training, was performed and showed that pretest problem type (plant, pendulum, or paint) was a significant covariate of pretest scores, $F(2, 126) = 4.199$, $p \leq .01$. No bias in the study was caused by this fact, however, as pretest problem type did not affect the overall posttest scores, $F(2, 126) = 1.077$, $p \leq .34$. Additionally, the type of problem given on the posttest was not significant, $F(1, 126) = .553$, $p \leq .4587$. The pretest scores did not predict posttest performance significantly either, $F(1, 126) = .215$, $p \leq .6439$.

Table 8

Mean Performance on the Posttest Using Pretest Performance and Posttest Problem Type as Covariates

Group	n	Pre mean	SD	Post mean	SD	Difference between means
Pretest–Posttest (G-1)	62	0.02	0.43	−0.08	0.40	−0.10
Intervention (G-2)	26	−0.10	0.40	0.16	0.46	0.26
Feedback (G-3)	9	0.40	0.52	0.05	0.43	−0.35
Reinforcement (G-4)	33	0.22	0.51	0.44	0.57	0.22
Construction (G-5)	4	0.20	0.37	−0.30	0.48	−0.50

Table 9
Posttest Analysis of Variance

Source of variance	SS	df	MS	F
Between subjects				
Covariates	0.17	2	0.08	0.38
Pretest	0.05	1	0.05	0.21
Posttest problem type	0.12	1	0.12	0.55
Group	6.41	4	1.63	7.30***

*** $p < .00005$.

Discussion

At the outset of the study, almost all of the participants performed at the Concrete Stage 9 or Abstract Stage 10 and not the Formal Stage 11. The participants were three to five years younger than those that typically exhibit Formal Stage 11 performance. By the end of the intervention, most of the participants whose correct answers were reinforced performed at Formal Stage 11 on the laundry problem (see Appendix) and in other types of problems. Although some participants gained Formal Stage 11 problem solving skills from practice and feedback alone, reinforcement was more successful in helping a greater number of the participants acquire this skill. The superiority of the reinforcement condition in bringing about change was shown in four different kinds of analyses, including the proportion of individuals in the groups who performed at the formal stage, the sensitivity of individuals in different groups to the causal relationship, across the trials of the experiment, the likelihood to show an increase toward formal operational performance across trials, and finally, the likelihood of performing at the formal operational level of a transfer of training task. These results suggest that there is a potentially important role for reinforcement to play in stage change and learning in general.

Several aspects of the study, that some could argue would bias the results, were found to have had no effect. This includes the type of posttest problem used. It should be noted that the aspect of the experimental design in which students were drawn from different years also did not add a bias to the study. Because some conditions and years were confounded, to make sure cohorts from the different years were equivalent, an analysis of variance was carried out on the Pretest–Posttest groups by year. No significant differences in detecting Formal Stage 11 relationships were found.

Differences in the initial performance of the participants could have biased the study as well. Although participants were randomly placed in groups before being given pretests, the mean performance of Reinforcement Group (G-4) participants was higher than Intervention Group (G-2) participants. Repeated-Measures ANOVA and the use of Trial 1 and 2 data as covariates statistically reduced the advantage that Reinforcement Group participants may have had over other groups by virtue of performing more formally at the beginning of the intervention. One might hypothesize that intervention alone would have a greater effect on participants that were closer to Formal Stage 11 performance. Some researchers (Cantor, 1983; Inhelder & Sinclair, 1969; Zimmerman & Blom, 1983) have suggested that participants whose performance is in transition between stages are more likely to demonstrate stage change in training studies than those whose performance is not. In the present study, the

Table 10
Post Hoc Comparison Between Groups on Posttest

Comparison	t test p <
Pretest–Postest (G-1) > Construction (G-5)	ns
Intervention (G-2) > Pretest-Posttest (G-1)	.0329
Intervention (G-2) > Feedback (G-3)	ns
Intervention (G-2) > Construction (G-5)	ns
Feedback (G-3) > Pretest-Posttest (G-1)	ns
Feedback (G-3) > Construction (G-5)	ns
Reinforcement (G-4) > Pretest-Posttest (G-1)	.0000
Reinforcement (G-4) > Intervention (G-2)	.0234
Reinforcement (G-4) > Feedback (G-3)	.0261
Reinforcement (G-4) > Construction (G-5)	.0033

Feedback Group (G-3) participants' mean initial level of performance was, in fact, higher than that of the Reinforcement Group (G-4). The Feedback Group (G-3) participants, however, showed no improvement across trials as Davidson (1983) found, whereas the Reinforcement Group (G-4) participants improved significantly. Thus, our results do not support this earlier suggestion that the varying initial performance of the participants caused a bias in the results of the study. After examining and discounting all of the possible biases the experimental methodology could have created, the overall result in all four data analyses that reinforcement is a sufficient condition for stage change to the Formal Stage 11 level remains valid.

Despite the efficacy of reinforcement, there are probably many educational experts and laymen alike who would object to this intervention technique because it might undermine the intrinsic or inculcated internal motivation that it seeks to compensate for. For example, Deci (1971, 1972) found that reinforcing events extrinsic to an activity may decrease the likelihood that participants will engage in that activity once these extrinsic reinforcers are removed. He attributed this phenomenon to a subsequent decrease in intrinsic motivation to engage in the task. Shiu Ling-po (cited by Hui, 2001) also maintains that "Award schemes are not something for teaching and learning. They may be able to motivate students at the beginning, but soon they will start asking their teachers for material benefits every time they are given an assignment." Some convincing evidence to the contrary of Ling-po and Deci's opinions that external reinforcement extinguishes internal motivation also exists. A study done by Flora and Flora (1999) provided no support for the assertion that extrinsic rewards for reading undermine intrinsic interest in reading. Instead, their study found that this external reinforcement set the conditions where intrinsic motivation for reading may develop. Although the long-term effects of external reinforcement are still being vigorously debated in academia and fall beyond the scope of the present study, reinforcement should not be completely dismissed as a way to help students acquire skills of higher complexity. Instead, because reinforcement could be such an effective teaching tool, as this study shows, investigations into its

long-term effects should be heightened. External reinforcement may also be particularly important for students from backgrounds in which arriving at a correct answer in and of itself is less important. The type of external reinforcement may also be important. Here, children could earn points for their team and this seemed to be particularly motivating for a number of children.

Another area that deserves further investigation is the effect of experience with the problem, or problem presentation, on the acquisition of higher stage problem solving skill. In this study all intervention groups that received more experience with the Formal Stage 11 problems performed significantly better than those who completed the Pretest-Posttest alone. As opposed to reinforcement, elapsed time might not be a factor sufficient in bringing about stage change to Formal Stage 11.

Some policymakers or researchers, like Deci (1971, 1972), have assumed that outcomes from doing teacher- or researcher-posed tasks are intrinsically reinforcing. It may be closer to the truth that those events that serve as reinforcers are determined culturally, socially, and by the interests of the individual. There may be participants for whom the consequence of trying to master the posed problems is reinforcing enough for transition to take place. Those participants who may not place as great of a value on merely getting a problem correct, however, may be able to acquire higher complexity problem solving skills through external reinforcement methods. Validating this possibility, the present study showed that the reinforcement intervention was the most effective method examined in getting the highest proportion of a highly diverse participant pool to accomplish Formal Stage 11.

This study was undertaken to get at the question of how stage change takes place in the real world. What may promote the acquisition of Formal Stage 11 in a restricted educational setting may not promote such change in everyday life. Also, there may be other methods that this study did not examine that are sufficient for inducing stage change as well. Nevertheless, it can be concluded that at least in an educational situation, the external reinforcement of correct answers along with practice and feedback helps more individuals develop problem solving skills

of a higher stage complexity than practice or feedback alone.

The Model of Hierarchical Complexity is contentless and domain free. It is also fractal so that the stage changes process is independent of stage. There is only one stage sequence and only one domain (Harrigan, Giri, & Commons, 2014). Therefore it is expected that these results will generalize to other content and other tasks and domains.

References

Bredderman, T. A. (1973). The effects of training on the development of the ability to control variables. *Journal of Research in Science Teaching, 10,* 189–200. http://dx.doi.org/10.1002/tea.3660100302

Brown, A. L., & Campione, J. C. (1990). Communities of learning and thinking, or a context by any other name. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills: Vol. 21. Contributions to human development* (pp. 108–125). Basel, Switzerland: Karger.

Brown, A. L., & Palincsar, A. S. (1989). Guided cooperative learning and individual knowledge acquisition. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 393–451). Mahwah, NJ: Erlbaum.

Cantor, G. N. (1983). Conflict, learning, and Piaget: Comments on Zimmerman and Blom's "Toward an empirical test of the role of cognitive conflict in learning." *Developmental Review, 3,* 39–53. http://dx.doi.org/10.1016/0273-2297(83)90006-0

Commons, M. L., Gane-McCalla, R., Barker, C. D., & Li, E. Y. (2014). The Model of Hierarchical Complexity as a measurement system. *Behavioral Development Bulletin, 19,* 9–14. http://dx.doi.org/10.1037/h0100583

Commons, M. L., & Goodman, S. A. (2008). Project Giant Step: Ocean Hill Brownsville and the New York City School strike. *Journal of Precision Teaching & Celeration, 24,* 38–42.

Commons, M. L., Kantrowitz, S., Buhlman, R. A., Ellis, J. B., & Grotzer, T. A. (1984). Immediately detecting a causal-relation stimulus embedded in similarly appearing possible causal relations. In M. L. Commons, M. C. Davison, J. A. Nevin, (Eds.), *Models of Behavior: Vol. 11, Signal Detection*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Commons, M. L., Li, E. L., Richardson, A. M., Gane-McCalla, R., Barker, C. D., & Tuladhar, C. T. (2014). Does the model of hierarchical complexity produce significant gaps between orders and are the orders equally spaced? *Journal of Applied Measurement, 15,* 422–449.

Commons, M. L., & Miller, P. M. (1998). A quantitative behavior-analytic theory of development.

Mexican Journal of Experimental Analysis of Behavior, 24(2), 153–180.

Commons, M. L., Miller, P. M., & Kuhn, D. (1982). The relation between formal operational reasoning and academic course selection and performance among college freshmen and sophomores. *Journal of Applied Developmental Psychology, 3,* 1–10. http://dx.doi.org/10.1016/0193-3973(82)90028-4

Commons, M. L., & Richards, F. A. (1980, September). *A developmental approach to tutoring.* Paper presented at the meeting of the American Psychological Association, Montreal.

Commons, M. L., & Richards, F. A. (1984a). A general model of stage theory. In M. L. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations: Vol. 1. Late adolescent and adult cognitive development* (pp. 120–140). New York, NY: Praeger.

Commons, M. L., & Richards, F. A. (1984b). Applying the general stage model. In M. L. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations: Vol. 1. Late adolescent and adult cognitive development* (pp. 141–157). New York, NY: Praeger.

Commons, M. L., & Rodriguez, J. A. (1987, January). *The lore of approaching tasks and the meaning of test lingo: Crosscultural interviews.* Paper presented at Universidad Autonoma de Baja California, Mexicali, Baja California, Mexico.

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). The existence of developmental stages as shown by the hierarchical complexity of tasks. *Developmental Review, 8*(3), 237–278.

Conrath, J. (1988). Dropout Prevention: Find out if your program passes or fails. *The Executive Educator, 10,* 15–16.

Crunch Software Corporation. (1987). *Crunch reference manual.* Oakland, CA: Author.

Davidson, M. N. (1983). *The effect of repeated presentation of a problem on the transition from concrete to formal operations: A test of Piaget's theory of stage change.* Unpublished B. A. honors thesis, Harvard University.

Davidson, M. N., & Commons, M. L. (1983, May). *The effect of repeated presentation of a formal operational problem with and without feedback on the transition from concrete through abstract to formal operations: An examination of Piaget's equilibration theory of stage change.* Paper presented at the meeting of the Jean Piaget Society, Philadelphia.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18,* 105–115. http://dx.doi.org/10.1037/h0030644

Deci, E. L. (1972). Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personal-*

ity and Social Psychology, 22, 113–120. http://dx .doi.org/10.1037/h0032355

Fargo, G. A., Behrns, C., Goodman, S. A., & Commons, M. L. (1971). Accountability: Utilizing behavior modification. In G. A. Fargo, C. Behrns, & P. Nolen (Eds.), *Behavior modification in the classroom.* Belmont, CA: Wadsworth.

Flavell, J. H. (1971). Comments on Beilin's "The development of physical concepts." In T. Mischel (Ed.), *Cognitive development and epistemology* (pp. 121–128). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-498640-4 .50009-3

Flora, D., & Flora, S. (1999). Effects of extrinsic reinforcement for reading during childhood on reported reading habits of college students. *The Psychological Record, 49,* 3–14.

Ginsburg, H. P., & Opper, S. (1988). *Piaget's theory of intellectual development: An introduction* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Grotzer, T. A., Commons, M. L., & Davidson, M. N. (1986, May). *The role of motivation in the mechanism of stage change: A consideration of equilibration theory and cultural values.* Presented at the meeting of the Jean Piaget Society, Philadelphia.

Grotzer, T. A., Robinson, T. L., Young, R. M., Davidson, M. N., Cohen, L., Rios, M. L., . . . Commons, M. L. (1985, June). *The role of repeated presentation of a formal operational problem, feedback, and social reward: An examination of Piaget's equilibration theory of stage transition.* Paper presented at the meeting of the Jean Piaget Society, Philadelphia.

Gruen, G. E. (1965). Experiences affecting the development of number conservation in children. *Child Development, 36,* 963–979. http://dx.doi .org/10.2307/1126937

Harrigan, W. J., Giri, S. P., & Commons, M. L. (2014). There is only one stage domain, put in the published information. *Behavior Development Bulletin, 19,* 51–61.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change.* New York, NY: Pergamon Press.

Hui, P. (2001, March 24). QEF projects come under fire. *South China Morning Post.* Retrieved from http://www.scmp.com/article/342384/qef-projects-come-under-fire

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York, NY: Basic Books, Inc. http://dx.doi .org/10.1037/10034-000

Inhelder, B., & Piaget, J. (1969). *The early growth of logic in the child.* New York, NY: Norton & Co.

Inhelder, B., & Sinclair, H. (1969). Learning cognitive structures. In P. Mussen, J. Langer, & M. Covington (Eds.), *Trends and issues in developmental psychology.* New York, NY: Rinehart & Winston.

Johnson, D. W., Maruyama, G., Johnson, R. W., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin, 89,* 47–62. http://dx.doi.org/ 10.1037/0033-2909.89.1.47

Kantrowitz, S., Buhlman, R. A., & Commons, M. L. (1985, April). *Application of signal-detection theory to the measurement of stage of development found in performance on formal-operational isolation-of-variables problems.* Paper presented at the meeting of the Eastern Psychological Association, Boston.

Kessen, W. (1971). Early cognitive development. In T. Mischel (Ed.), *Cognitive development and epistemology* (pp. 287–307). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-498640-4.50014-7

Kitchener, K. S., & Fischer, K. W., (1990). A skill approach to the development of reflective thinking. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills: Contributions to Human Development* (Vol. 21, pp. 48–61). Basel, Switzerland: Karger.

Kuhn, D. (Ed.), (1990). *Developmental perspectives on teaching and learning thinking skills: Contributions to Human Development* (Vol. 21). Basel, Switzerland: Karger.

Kuhn, D., & Angelev, J. (1976). An experimental study of the development of Formal Stage 11 thought. *Child Development, 47,* 697–706. http:// dx.doi.org/10.2307/1128184

Kuhn, D., & Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. *Developmental Psychology, 13,* 9–14. http://dx.doi.org/10.1037/ 0012-1649.13.1.9

Kuhn, D., & Ho, V. (1977). The development of schemes for recognizing additive and alternative effects in a "natural experiment" context. *Developmental Psychology, 13,* 515–516. http://dx.doi .org/10.1037/0012-1649.13.5.515

Kuhn, D., Ho, V., & Adams, C. (1979). Formal reasoning among pre- and late-adolescents. *Child Development, 50,* 1128–1135. http://dx.doi.org/ 10.2307/1129340

Langer, J. (1969). Disequilibrium as a source of development. In P. Mussen, J. Langer, & M. Covington (Eds.), *Trends and issues in developmental psychology.* New York, NY: Holt, Rhinehart and Winston.

Linn, M. C., Chen, B., & Thier, H. D. (1976). Personalization in science: Preliminary investigation at the middle school level. *Instructional Science, 5,* 227–251. http://dx.doi.org/10.1007/BF00054357

McCann, D. C., & Prentice, N. M. (1981). Promoting moral judgment of elementary school children: The influence of direct reinforcement and cognitive disequilibrium. *The Journal of Genetic Psychology: Research and Theory on Human Development, 139,* 27–34. http://dx.doi.org/10.1080/00221325.1981.10533433

Miller, R. G., Jr. (1981). *Simultaneous statistical inference* (2nd ed.). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4613-8122-8

Munsinger, H. (1970). Light detection and pattern recognition: Some comments on the growth of visual sensation perception. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology* (pp. 227–247). New York, NY: Academic. http://dx.doi.org/10.1016/B978-0-12-293850-4.50015-7

Neimark, E. (1975). Intellectual development during adolescence. In F. Horowitz (Ed.), *Review of child development research* (Vol. 4). Chicago, IL: University of Chicago Press.

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. Chicago, IL: The University of Chicago Press.

Piaget, J. (1965). *The child's conception of number*. New York, NY: W. W. Norton and Company.

Richard, D. C., Unger, C. M., & Commons, M. L. (April, 1988). *Strategies as knowledge: Participants' methods of determining causality during the shift from concrete to formal operations*. Paper presented at the meeting of the Eastern Psychological Association, Buffalo, NY.

Rodriguez, J. A., Buhlman, R. A., Kantrowitz, S., & Commons, M. L. (1986, May). *A signal-detection analysis of individual Nigerian children's explanatory statements on the formal "plant" problem showed very little stage mixing*. Paper presented at

the meeting of the Jean Piaget Society, Philadelphia.

Rosenthal, R., & Rosnow, R. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, UK: Cambridge University Press.

Siegler, R. S., & Atlas, M. (1976). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Detecting interactive patterns in data. *Journal of Educational Psychology, 68,* 360–370. http://dx.doi.org/10.1037/0022-0663.68.3.360

Siegler, R. S., Liebert, D. E., & Liebert, R. M. (1973). Inhelder and Piaget's pendulum problem: Teaching preadolescents to act as scientists. *Developmental Psychology, 9,* 97–101. http://dx.doi.org/10.1037/h0035073

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York, NY: Appleton.

Swets, J. A., & Green, D. M. (1961). Sequential observations by human observers of signals in noise. In C. Cherry (Ed.), *Information theory*. London, UK: Betterworth.

Swets, J., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340. http://dx.doi.org/10.1037/h0040547

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1966). Development of the higher mental function. In A. Luriya & A. Smirnov (Eds.), *Psychological research in the U.S.S.R.* (pp. 44–45). Moscow, Russia: Progress Publishers.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Press.

Zimmerman, B. J., & Blom, D. E. (1983). On resolving conflicting views of cognitive conflict. *Developmental Review, 3,* 62–72. http://dx.doi.org/10.1016/0273-2297(83)90008-4

(*Appendix follows*)

Appendix

Laundry problem layout

Information Episodes for Trial 1. The cloth was stained with red lipstick. It was washed in each of these six ways.

A Bleach Powder Soap Blue Booster Cold Water ⟶ Dirty

B Bleach Liquid Soap Pink Booster Hot Water ⟶ Clean

A Bleach Powder Soap Pink Booster Hot Water ⟶ Dirty

B Bleach Powder Soap Pink Booster Cold Water ⟶ Dirty

A Bleach Liquid Soap Blue Booster Hot Water ⟶ Clean

B Bleach Liquid Soap Blue Booster Cold Water ⟶ Clean

Note: The bottles of washing ingredients are laid out in front of the subject vertically, so that (in this trial) the bleach is in the back row and the water in the front row. The clean or dirty cloth is always the nearest object to the subject

Prediction Episodes for Trial 2
Look back at the examples. Now, mark the correct ending.

B Bleach Powder Soap Blue Booster Hot Water ⟶ Clean   Dirty

A Bleach Liquid Soap Blue Booster Cold Water ⟶ Clean   Dirty

A Bleach Powder Soap Pink Booster Cold Water ⟶ Clean   Dirty

B Bleach Liquid Soap Blue Booster Hot Water ⟶ Clean   Dirty

B Bleach Powder Soap Blue Booster Cold Water ⟶ Clean   Dirty

B Bleach Powder Soap Pink Booster Hot Water ⟶ Clean   Dirty

A Bleach Liquid Soap Pink Booster Hot Water ⟶ Clean   Dirty

A Bleach Powder Soap Blue Booster Hot Water ⟶ Clean   Dirty

B Bleach Liquid Soap Pink Booster Cold Water⟶ Clean   Dirty

A Bleach Liquid Soap Pink Booster Cold Water ⟶ Clean   Dirty

Note: One episode is represented to the subject at a time, with the bottles arranged vertically and no cloth in front of them.